音声利用向けに発音情報を出力できるタイ語解析器の開発

山崎 智弘　　宮村 祐一　　山中 紀子
東芝研究開発センター 知識メディアラボラトリー
{tomohiro2.yamasaki, yuichi.miyamura, noriko.yamanaka}@toshiba.co.jp

1 はじめに

ここ数年の円安傾向や政府による訪日旅行促進事業により日本を訪れる外国人の数は増加傾向にあるが、「東京オリンピック開催の 2020 年に 2000 万人」という政府目標が先日 3000 万人に引き上げられたように、今後も大幅に増加することが見込まれる。

図 1 [3] からわかるように、日本を訪れる外国人の数が多いのはアジア諸国である。その中で東南アジア諸国は、韓国・中国・台湾・香港と比べると絶対数はまだそれほど多くないが、伸び率が大きいことや人口が多いため伸びしろが大きいことを考慮すると、インバウンド事業における重要性が一層高まるであろう。特にタイ語・インドネシア語・ベトナム語・ミャンマー語は総務省によるグローバルコミュニケーション計画で挙げられていることもあり、我々は近年、インバウンド事業の基盤技術としてこれらの言語でも音声対話・音声翻訳を実現するべく研究開発を進めている。

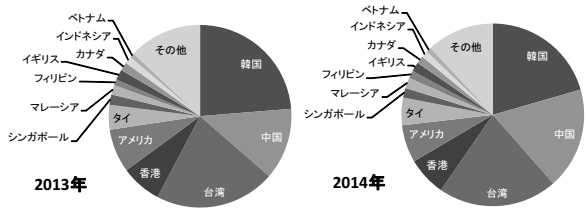

図 1: 日本を訪れた外国人の国別割合

一般に、適切な情報が揃った単語セット (辞書) と適切なメタデータが付与されたテキスト (コーパス) が大量にあれば、言語解析器は機械学習によって比較的容易に作れることが知られている [5]。すなわち言語知識としては、詳細な文法の記述よりも

- 表記・品詞・発音を網羅した単語セットの収集
- テキストへの有用なメタデータの高精度な付与

に重点が置かれることになる。

しかし東南アジア言語の商用利用可能な辞書やコーパスは日英中韓と比べると圧倒的に手に入りにくい。また発音情報を出力できるようにするには単語の発音まで必要であるが、ネイティブには不要なのか音声利用までは考えていないのか、発音まで揃った辞書はさらに手に入りにくい。品詞についても、ヨーロッパ言語 (英語など) の体系を流用していて東南アジア言語の解析には適切な体系になっていないことも多い。

そこで我々は、東南アジア言語の中からまずタイ語を取り上げ、テキストにメタデータをリアルタイムで付与・編集できるシステムの開発、および網羅的に単語の正解発音を収集する手順の策定を行ない、自前で辞書とコーパスを整備した。その後それらを元に機械学習し、入力された文の発音情報を出力できるタイ語解析器を開発した。タイ語は日本語と同じく分かち書きしないため、本解析器は入力された文を単語分割したあと品詞推定・発音推定するものとなっている。

本論文では簡単にタイ語の特徴を述べた後、まずメタデータ編集システムとコーパス整備手順の概要をまとめる。続いて機械学習でどのようにタイ語の単語分割・品詞推定しているか説明し、評価を行なう。その後、タイ語の表記から発音を推定する手法およびそれを用いた正解発音収集手順をまとめ、発音推定の評価を行なう。最後にタイ語以外の東南アジア言語にも考察を加え、今後の課題についてまとめる。

1.1 タイ語の特徴

本節では簡単にタイ語の特徴について述べる。

- 表記は外来語も含めインド系表音文字のタイ文字で分かち書きせずに書く。句読点を用いない。
- 発音は頭子音+母音+末子音+声調からなる音節で構成される。声調は平・上・高・下・低の 5 つ。
- 語形変化せず 1 つの単語が多くの品詞で使われるため基本語は少ない (= 複合語や句が多い)。

表音文字なので発音は基本的に表記から定まるが、規則は複雑かつ例外も多い。特に文字の組合せによっ

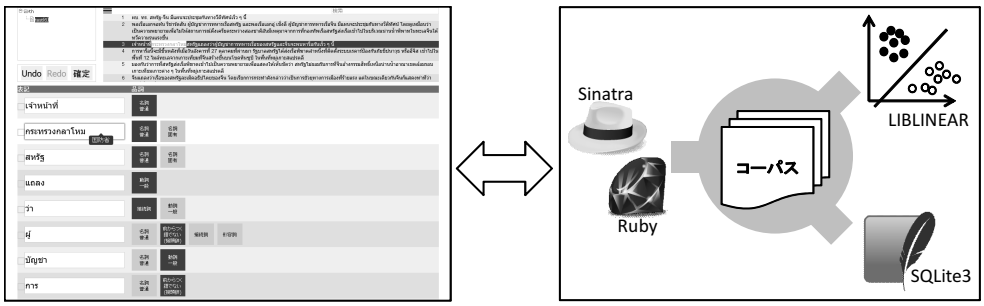

図 2: アーキテクチャ概要

ては複数の規則が当てはまるため、単語や音節の境界がわからないと文を正しく発音できない。しかし複合語や句でも単語の発音が変化しないので、文の発音推定は文の単語分割と単語の発音推定に還元できる (= 単語の発音をつなげるだけで文をほぼ正しく発音できる) という特徴もある。

2 単語分割と品詞推定

単語分割と品詞推定には SVM を用いた点予測手法 [4] を採用した。本手法は周囲の単語境界や品詞を素性として用いなくてよいため、コーパスの整備コストを抑えることができるという特徴がある。

コーパス整備ではテキストファイルをエディタで編集するということがよく行なわれるが、分割を修正したのに品詞を修正し忘れる、修正するべき個所を見逃す、誤った品詞を付与したのに気づかないといったミスが起こりやすい。しかし修正するたびにその修正結果に基づいて推定し直せば、依存関係のあるデータ (分割と品詞、品詞の並びなど) も適切に修正されるはずであるし、ミスの可能性が高いときは警告を出すことができるはずである。

そこで我々はバックエンドに学習器と分類器を組み込み、テキストにメタデータをリアルタイムで付与・編集できるだけでなく、修正したコーパスを元に機械学習することでモデルを随時改善していくことができるメタデータ編集システムを開発した。次節ではメタデータ編集システムの概要について説明する。

2.1 メタデータ編集システム

我々が開発したメタデータ編集システムはサーバ･クライアント型のウェブアプリケーションである。サーバは Ruby (Sinatra) で記述されている。データの格納には SQLite3、学習器と分類器には LIBLINEAR を用いる。クライアントは JavaScript (jQuery) で記述されており、サーバと通信してメタデータ付与の対象となる文や学習済みモデルに基づく解析結果を取得するほか、修正されたメタデータの送信とそれに基づく新たな解析結果の取得を行なう。

メタデータ付与の作業単位としては「文」が望ましいが、1.1 節で述べたようにタイ語には句読点がないため、機械的に文に分割しておくことが難しい。とはいえ「段落」は非常に長くなることがあり、作業単位として扱いづらい。そこで段落内の意味的な切れ目に挿入されているスペースを便宜的に句読点と見なし、その前後で分割して作業単位とするものとした。

品詞体系としては、文内での単語の働きを決めるのに十分な区別があるだけでなく、作業者が明確に区別できなければならないという観点から、表 1 に示す 18 種類を定義した。

表 1: タイ語の品詞体系

大分類	中分類	説明
名詞	一般	一般名詞
	固有	固有名詞
	類別	類別詞 (日本語でいう助数詞)
動詞	一般	一般動詞
	補助	補助動詞 (日本語でいう助動詞)
形容詞	−	形容詞、日本語でいう副詞の一部
副詞	−	文を修飾する副詞
接辞	接尾	必ず何かの後につき単独で使わないもの
	接頭	必ず何かの前につき単独で使わないもの
接続詞	−	接続詞 (並立だけでなく従属も)
感嘆詞	−	感嘆詞
助詞	前置	英語でいう前置詞
	後置	日本語でいう格助詞
	文末	日本語でいう終助詞
その他	句読	句読点 (タイ語のものはない)
	記号	句読点以外の記号
	数字	数字 (タイ数字も)
	英字	英字

その後、タイのニュースサイト [1] から収集したニュース記事に対し、本システムでタイ語に堪能な作業者 2 名にメタデータを付与してもらい、22 万語弱からなる品詞付きコーパスを整備した。

2.2 単語分割と品詞推定の評価

本節では 2.1 節で述べたコーパスから学習用に約 700 段落 20 万語、評価用に約 100 段落 2 万語を抽出して行なった実験について述べる。

単語分割は各文字境界について単語境界かどうか推定する 2 値分類として定式化されるため、各文字境界の前後 3 文字以内の文字 N-gram (N = 1, 2, 3) を素性として学習・分類した。各文字境界について単語境界かどうか正しく判定できた割合は 0.990 であった。単語分割は単語内のすべての文字境界で正しく判定できなければならないため実際の単語分割精度はもう少し低くなるが、単純な素性ながら非常に高い精度が出ていることがわかる。

品詞推定は表 1 にある 18 値分類として定式化される。単語分割と同じく単語周辺の文字 N-gram を素性としてもよいが、前後の品詞 (またはそれの代替としての表記) の方が有効と考え、各単語の前後 3 単語以内の単語 N-gram (N = 1, 2, 3) を素性として学習・分類した。単語分割が正しく行なわれたという仮定のもとで品詞が正しく推定できた割合は 0.909 であった。正しく推定できなかった事例を調査したところ、普通名詞を固有名詞や類別詞、接続詞を前置詞、形容詞を副詞に間違えやすいことがわかった。これらは作業揺れが生じやすい事例でもあり、精度向上のためには作業揺れを解消してコーパスの精度を高める必要があると考えられる。

3 発音推定

1.1 節で述べたように、タイ語の発音は声調も含めて基本的に表記から定まる。

日本語と同じく数字は、小数点以下は文字読みすればよいが一般には桁読みする必要がある。算用数字のほかに表 2 のようなタイ数字というものがあるが、タイ語は語形変化しないので、桁読みは桁の値の単語と桁の単語を単純に並べればよい。

表 2: タイ数字と桁の値の単語、および桁の単語

๐	๑	๒	๓	๔	๕	๖	๗	๘	๙
0	1	2	3	4	5	6	7	8	9
ศูนย์	หนึ่ง	สอง	สาม	สี่	ห้า	หก	เจ็ด	แปด	เก้า

10^1	10^2	10^3
สิบ	ร้อย	พัน
10^4	10^5	10^6
หมื่น	แสน	ล้าน

数字でない単語は、頭子音+母音+末子音+声調からなる音節を特定する必要がある。頭子音・母音・末子音は表記と発音がほぼ一対一に対応しているが、声調を定める規則は複雑である。また一部の母音は子音としても使われたり表記されなかったりするため、音節の境界が一意に特定できないことも多い。1 つの子音を末子音・頭子音として 2 回読んだり特定の子音を読まなかったりする例外的な単語も比較的多いため、正しい発音情報を出力するためには発音まで揃った辞書が必要となる。

3.1 正解発音の収集手順

例外的な発音の単語を除くと、音節を特定できれば表記から発音を推定できると期待される。タイ語における音節分割は日本語における単語分割と同じ問題と見なせるため、最も単純な文節数最少法を流用して発音候補を推定するものとした。具体的には、規則に基づいて表記と音節の組合せを網羅しておき、与えられた表記に対して音節ノードからなるラティスを生成し、最短経路を求めることで発音候補を推定する。

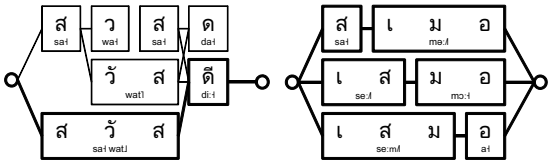

図 3: 音節ラティスの例 (左: สวัสดี、右: เสมอ)

本手法ではタイ語としてありえない表記でないかぎり何らかの発音候補が得られるが、あくまでも発音候補でしかないことに注意が必要である。例えば図 3 左の สวัสดี は正解発音が得られているが、右の เสมอ は発音候補が複数ある。また 1 つしかなくても一般には正解発音とはかぎらない。正解発音を特定するためには、信頼できるソースとの比較やタイ語に堪能な作業者による確認が必要である。

信頼できるソースとしては Wikipedia が考えられたが、ほとんどの見出しに発音が記載されておらず活用できなかった。しかし Wiktionary [2] という姉妹プロジェクトがあり、そちらは発音が記載された見出しが多いことがわかった。ただしタイ語独自の表記法で記載されていることが多かったため、IPA (国際音声記号) や X-SAMPA (拡張 SAM 音声記号) へ変換しつつ比較した。

さらにタイ語 Wiktionary には単語や音節の境界が記載された見出しも多いことがわかった。そこで、それらの境界ごとに発音候補が一意に得られたときはそれらをつなげて全体の正解発音とした。Wiktionary との比較で正解発音が特定できなかった単語については、タイ語に堪能な作業者 2 名に正解かどうか判定してもらった。

これらをまとめると以下のような手順となる。

1. Wiktionary ダンプデータから発音を抽出する。推定した発音候補と抽出結果が一致したときはそれを正解と見なす。
2. Wiktionary ダンプデータから境界 (単語・音節) を抽出する。境界 (単語・音節) に基づいて発音候補が一意に得られたときはそれを正解と見なす。すでに正解発音が特定できた単語で発音候補が一意に得られたときもそれを正解と見なす。
3. タイ語に堪能な作業者に発音候補の○×を判定してもらう。出現頻度の高い単語の発音候補を用意し、○と判定されたときはそれを正解と見なす。

3.2 発音推定の評価

上述の手順に従って正解発音を収集する実験を行なったところ、20282 語の正解発音が特定でき、発音候補に実際に正解発音が含まれる割合は 0.817 であった。具体的には、推定した発音候補と抽出結果が一致したのは 12043 語、境界 (単語・音節) に基づいて発音候補が一意に得られたのは 1619 語であった。適当な出現頻度で足切りを行なって得た頻出語のリストからこれらの単語を除いた 8022 語のうち、すでに正解発音が特定できた単語で発音候補が一意に得られたのは 352 語、タイ語に堪能な作業者 2 名に判定してもらって正解発音が特定できたのは 6268 語であった。

一方 1.1 節で述べたように、文の発音推定は文の単語分割と単語の発音推定に還元されるので、文内の単語のうち正解発音が付与されているものの割合で文の発音推定の精度を検証した。表 3 は 2.1 節で述べたコーパスに出現した単語のうち発音が付与されている・いないものの延べ数である。

3 章で述べたように、タイ語は数字の発音が規則的なので正しく発音情報が出力される割合は概ね $(153375+53999)/216866 = 0.953$ となる。ただし数字などは正解発音の収集対象でないため、収集対象の単語のうち発音が付与されている割合は $153375/(153375+9492) = 0.942$ となる。

表 3: 発音が付与されている・いない延べ単語数

付与されている	153375
付与されていない	9492
その他 (数字など)	53999
計	216866

なお発音が付与されていない 9492 語の異なり数は 3866 語であり、10 回以上出現したものは 129 語であった。これらの事例を調査したところ、表 4 に示すように英語由来か固有名詞 (人名・地名、製品名など) が多いことがわかった。また発音が付与されていない単語であっても発音候補に正解があることが多いため、発音候補からうまく正解が選べれば未登録語に対する発音推定のロバスト性が高まると考えられる。

表 4: 収集漏れで発音が付与されていない単語の例

โอ.เค. OK ○ o:˧ k^{h}e:˧	โอบามา オバマ ○ o:˧ ba:˧ ma:˧
เรดาห์ レーダー × ra˧ daw˧ ○ re:˧ da:˧	ปิโตรเลียม 石油 × piɴ to:n˧ liam˧ ○ piɴ tro:˧ liam˧

4 おわりに

本論文では東南アジア言語の中からタイ語を取り上げ、我々が自前で整備した辞書とコーパス、およびそれらを元に開発した「入力された文の発音情報を出力できるタイ語解析器」の処理について説明した。

文字 N-gram の点予測による単語境界推定は 0.990、単語 N-gram の点予測による品詞推定は 0.909 の精度であった。また発音推定については、発音候補に正解発音が含まれる割合は 0.817、正解発音が特定できた単語の網羅率は 0.942 であった。

ただし本論文で述べた正解発音の収集手順は、発音候補に正解発音が含まれない単語はケアしていないため、そのような単語の正解発音を効率よく入力するためのＵＩは検討が必要である。

とはいえ東芝が持つ日本語・英語・中国語・韓国語という音声向け言語解析技術のラインナップにタイ語が加わったことになる。ベトナム語・インドネシア語・ミャンマー語は、タイ語と同じくほとんど語形変化しない言語であるため、今回の手法はこれらの言語にも適用できると考えられる。インドネシア語と特にミャンマー語は複合語や句で発音が変化するのでその点に留意する必要はあるが、開発を進めて東南アジア言語のラインナップを増やしていく予定である。

参考文献

[1] http://www.komchadluek.net, http://www.bangkokbiznews.com, http://www.manager.co.th.

[2] https://www.wiktionary.org. Wiktionary. Wikimedia Foundation.

[3] 日本政府観光局 (JNTO). 国籍/月別 訪日外客数, 2003--2015. http://www.jnto.go.jp.

[4] Graham Neubig, 中田陽介, 森信介. 点推定と能動学習を用いた自動単語分割器の分野適応, 2010 年 3 月. 言語処理学会第 16 回年次大会 (NLP2010).

[5] 山崎智弘, 若木裕美, 清水勇詞, 鈴木優. 統計的手法に基づく品詞解析器の半自動構築, 2011 年 2 月. 第 3 回データ工学と情報マネジメントに関するフォーラム (DEIM2011).